DOCUMENT RESUME

ED 395 998 TM 025 218

AUTHOR Ryan, Katherine E.; Chiu, Shuwan
TITLE Detecting DIF on Mathematics Items: The Case for Gender and Calculator Sensitivity.
PUB DATE Apr 96
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Calculators; *College Students; Content Analysis; Equal Education; Higher Education; *Item Bias; *Mathematics; Regression (Statistics); Sample Size; *Sex Differences; Standardized Tests; *Test Items
IDENTIFIERS *Item Bias Detection; Rule Space Model

ABSTRACT

The use of scientific calculators on standardized mathematics tests is becoming more common, and how their use affects the standardized testing process continues to be a topic of study. This study extends previous work by examining test item characteristics and equity issues, asking whether items designed to be "calculator neutral" function as intended, and whether such items function the same for males and females. Data came from trials of the Midwestern Mathematics Placement Examination for precalculus and calculus classes for over 1,000 college students. A logistic regression analysis was conducted to detect differential item functioning (DIF) or nonuniform DIF. Findings suggest that calculator neutral items can be constructed, and that the items that were constructed to be calculator neutral were largely free from gender DIF. Results also showed that the logistic regression approach was a useful addition to studying DIF, although larger sample sizes would have been highly preferable. Using the rule-space approach as a basis for the content analyses was also quite effective. An appendix contains two sample problems and a description of the mathematical challenge items. (Contains 4 tables, 4 figures, and 17 references.) (SLD)

DETECTING DIF ON MATHEMATICS ITEMS: THE CASE FOR GENDER AND CALCULATOR SENSITIVITY

Katherine E Ryan
Shuwan Chiu
Division of Measurement and Evaluation
University of Illinois
182 Armory Building
505 East Armory
Urbana, IL 61801

For presentation at the Annual Meeting of the American Educational Research Association, New York, NY, April, 1996

Detecting DIF on mathematics items: The case for gender and calculator sensitivity

INTRODUCTION

The use of scientific calculators on standardized mathematics tests is becoming more common since the National Council of Teachers of Mathematics' recommendation on this issue. (NCTM, 1989). For example, both the ATP Mathematics Level II Achievement Test and the mathematics portion of the Scholastic Assessment Test (SAT), published by Educational Testing Service, permit testtakers to use calculators when answering the test items. Consequently how the use of calculators impacts the standardized testing process continues to be a topic of investigations. Issues studied include effects on test speededness (Ansley, Spratt, & Forsyth, 1989; Harvey, Jackson & Facher, 1993; Loyd, 1991;) and sensitivity of items to calculator use (Cohen & Kim, 1992, Harvey, et al.)

The results of research on test speededness is mixed; minimal and moderate speededness effects were found (Ansley, et al., 1989; Bridgman, Harvey, & Braswell, 1995; Harvey, et al., Loyd, 1991). The findings from studies on items classified as inactive (no advantage or disadvantage in calculator use), neutral (item can be solved with or without a calculator), or active (a calculator is necessary/helpful in answering the questions) with respect to item sensitivity to calculator use is more clear-cut. Examinees who used calculators had an advantage on neutral and active items while calculators were used infrequently on inactive items (Harvey, et al., 1993) In a study by Cohen and Kim, of the 28 items on the test, 5-12 items were susceptible to calculator effects depending on the method used (subscore and item level analyses)

Other areas of investigation include the effects of calculator use on item and test characteristics and gender differences (Ansley et al., 1989, Cohen & Kim, 1992, Harvey et al.,

Loyd, 1991). Cohen and Kim found no significant difference in an analysis of the test scores between the use/non use of calculators on a college placement exam; however, several calculator effects were detected in item level analyses In a study involving high school students, no advantage to calculator use was detected even though 19 out of 25 items required low-level computation (Ansley et al., 1989). Loyd (1991) found a decrease in coefficient alpha items administered with calculators when comparing the performance of items answered with and without calculators.

Using an ANCOVA design (controlling on achievement), Bridgman et al. found males and females benefited equally from calculator use at the test score level. However, in the Harvey et al investigation, the Mantel-Haenzel procedure was used to detect differential item functioning for males and females. Females found all types of calculator items (active, inactive, neutral) differentially more difficult than males The mean MH-D-DIF values by item type ranged from moderate ($\overline{X}$=-.25; SD=.59) for calculator inactive items to substantial ($\overline{X}$=-.48; SD=.50) for calculator neutral items

This study seeks to extend the work by Harvey et al., by further examining item characteristics and equity issues. The investigation is designed to address two questions. First, do items designed to be calculator-neutral, (i.e the type of item where calculator use might be helpful, but the item can be solved without a calculator use), function as intended? Second, do calculator neutral items function the same for males and females?

METHODS

Tests

The Midwestern Mathematics Placement Exam (MMPE) pilot items are based on course content covered in pre-calculus colleg : courses. While all in-coming freshman with three years of high school mathematics will be administered the test, the purpose of the test is to place students in a pre-calculus course and a first semester calculus course. The test is a 'low stakes assessment.' Students are not required to follow course placement recommendations based on the MMPE test score results. Nevertheless, accurate course placement is useful and efficient for students, faculty, and the institution (Ryan & Fan, 1993). Fairness is also a concern; particularly in light of recent research which suggests that female performance in college mathematics courses is underpredicted by college entrance examinations like the Scholastic Aptitude Test-Mathematics (SAT-M) (Linn & Kessel, 1995; Wainer & Steinberg, 1993).

The previous version of the MMPE was adequate for placing the students in appropriate courses; however, the exam resembled an aptitude test and did not reflect modern instructional approaches to mathematics like calculator use. Consequently, the pilot test was designed to allow calculator use and was composed of algebra, trigonometry, and geometry items. However, to avoid some of the standardization issues surrounding calculator use in tests like equal access to calculators, test items were designed to be primarily calculator neutral.

Four-six items were randomly assigned to two pilot test forms, each with twenty-three items: Forms A and B. The test instructions for Forms A and B indicated no calculator use was allowed. Two other test forms were assembled: Forms C and D. These forms were identical to Form A and Form B, respectively. However, the test instructions for Forms C and D permitted ordinary scientific calculator use when answering the test items. Test instructions indicated students were allowed 40 minutes to complete the test.

Design and Sample:

Data used in this investigation were collected from two item trials: Spring, 1995 and Fall 1995. The test forms were administered in pre-calculus and calculus classes. For both item trials, test booklets were spiraled to create equivalent groups for data collection. The Spring 1995 study sample consisted of 346 undergraduates in pre-calculus and introductory calculus courses at a large midwestern university. The sample size for the test forms ranged from 82 students completing Form D to 94 students completing Form A.

Any item which was not functioning as intended according to the content and or statistical review was either deleted, revised, or re-classified. The Fall study involved a large scale administration of the revised test items which took place on August 25, 1995. Over one thousand testtakers in pre-calculus and calculus courses participated with the number of examinees completing each form ranging from 249 for Form B to 316 for Form A. Item data used in this study are from items that were common to both item pilots and were functioning as intended. Consequently seventeen items were retained from A and C and twelve items from B and D were retained for further analyses.

Analyses:

A logistic regression analysis was conducted for each test item to detect uniform DIF (after controlling on achievement, the probability of answering the item under study is greater for one group in comparison to another) or non-uniform DIF (the difference in answering the item correctly for matched groups of test takers is not the same for all achievement levels) (Swaminathan & Rogers, 1990). Logistic regression has the following formulation:

$$P(u = 1) = \frac{e^z}{(1 + e^z)}$$

$$z_i = \tau_0 + \tau_1 G_i + \tau_2 X_i + \tau_3(X_i g_i),$$

i is the index for the testtaker; u_i is the testtaker's item response scored a 1 (correct) or 0 (incorrect), X_i is the testtaker's total score, and G_i is the testtakers group; $G = 1$ if the examinee is a member of the focal group; 2 if the examinee is a member of the reference group.

The SAS-PC Proc Catmod procedure was used for each comparison (e.g., calculator/no calculator) to estimate the parameters, τ_1, τ_2, τ_3. The dependent variable, test item response, was coded as 0 (incorrect) or 1 (correct). The total score on the MMPE pilot items (X, achievement) was designated as the continuous independent variable or covariate. Members of the reference group (standard to compare performances of the focal group) were assigned a '2'; focal group members (subgroup of interest) were assigned a '1' for the group membership variable. If the sign of the estimate of the parameters (τ_1, or τ_3) is positive, the focal group is favored; otherwise. the reference group is favored. Models were evaluated with a chi-square statistic with 1 degree of freedom. Calculations for logistic regression are described in detail elsewhere (Swaminathan & Rogers, 1990).

<u>Analyses Design</u>

Two basic models were tested sequentially with a backwards procedure to examine gender, calculator use (CNC) and achievement effects. Table 1 provides a description of the logistic regression analyses that were conducted To test for non-uniform DIF, the parameter for a three-way interaction model was tested (τ_3) (Set 1) (See Camilli & Shepard, 1994 for details of

this formulation). To examine the effects of calculator use, the model included an interaction term for calculator use by right/wrong by achievement. To investigate gender effects, the interaction term for gender by right/wrong by achievement was tested. To detect uniform DIF, the items free from non-uniform DIF were examined with a simpler model. With this model. the parameter, τ_1 is tested. (Set 2). This model consisted of total score, calculator use, and gender as independent variables; the gender and calculator use parameters for each item were inspected.

Each of the 17 items were tested for Forms A and C and 12 items were tested for Forms B and D; the criterion was total test score on test items for all analyses The studied item was included in the criterion Test takers completing Form C (calculators) were specified as a reference group; examinees answering Form A (no calculators) were designated as the focal group To investigate gender effects, the corresponding forms were combined (e.g. test forms A and C), males from A and males from C served as the reference group; females from test forms A and C were specified as the focal group. Parallel analyses were conducted with the examinees' responses to Forms D and B of the test to replicate findings from Forms C and A.

----Insert Table 1----

Content analyses were conducted for any DIF items identified. Traditionally, content analyses of DIF items are based on Bloom's taxonomy or inspection (Nandakamur, 1993; Ryan & Fan, 1994) Instead, the content analyses are based on the Rule-Space approach developed by k. Tatsuoka (1993). This approach was adapted for reporting the math proficiencies for the new SAT-M (Harnisch, Tatsuoka, & Wilkins, 1995) Items are inspected in relationship to a set of attributes which are the cognitive kills necessary to answer the test question correctly. (See Appendix A for a list of attributes called math challenges.)

RESULTS

The tests were designed to be parallel in content and difficulty. A summary of the descriptive statistics for the total sample, by form and by gender is reported in Table 2. The A and C (AC) combined version of the test is somewhat easier than the B and D (BD) combined test form ($\overline{X}$=7.16 for 17 items versus $\overline{X}$=4.14 for twelve items). There are minimal differences between the corresponding calculator and non-calculator forms of the test (less than .15 s.d. for A-C and B-D forms). There are modest differences in test performance between males and females (less than 4 s.d. for the AC BD forms).

----Insert Table 2-----

Table 3 presents the results for the reliability analyses. The coefficient alpha estimates for the matching calculator and no-calculator forms were approximately .65 for Form AC and .61 for Form BD. However when the estimates were calculated separately, the reliability estimates for the forms which allowed calculator use (Forms C and D) were slightly higher (.63 vs .67 and .58 vs 64). However these differences are not statistically significant (z = .24 for Forms A and C; z = .26 for Forms B and D ($p > .05$)) Estimates for the Spearman-Brown prophecy formula were also calculated for 40 items and 45 items. The estimates ranged from .87 for Form D to .84 for Form C. The differences in the Spearman-Brown reliability estimates for the calculator and non-calculator versions (40 and 45 items) of the tests were not statistically significant (not reported).

----Insert Table 3----

The results for the logistic regression analyses are presented in Table 4. The parameter estimate for the three-way interaction term was not statistically significant for any items from the test forms However, uniform DIF was detected for 4 items (Set 2 Analyses) Two items were

found to be differentially functioning when the items were tested using the simpler model for Forms A and C. The gender main effect was statistically significant for items 7 and 9; these items are differentially easier for male test takers

----Insert Table 4----

Plots of the score distributions and the probability of a correct response for male and female test takers for items 7 and 9 are presented in Figures 1 and 2. As shown in the Figures 1 and 2, the probability of a correct response on this items is not the same for males and female testtakers at the same achievement levels. For example, for students who scored a 5 or 6 on item 9, the probability of a correct response for men to answer the item correctly is approximately .6. In contrast, the probability of women (who scored a 5 or 6 on item 9) answering this question is around .43.

Based on an analysis of the attributes (attributes 1,2, 3, 12), question 7 involves a function with a second degree algebraic expression. However, if testtakers did not know how to solve the function, they could use a test-taking strategy and work backwards. They can compare the values to find the answer. The results from this item suggests women may be weaker in test-taking skills. (See Appendix A for the text of item #7.) In order to solve the other problem (9), testtakers need to know how to translate word problems into an algebraic expression and restructure the problem into a solvable form (attributes 5 and 6).

--Insert Figures 1 and 2 here----

Item 1 on Forms B and D is also differentially more difficult for females. (See Figure 3). In addition the calculator version of item 2 is differentially easier for test takers. Figure 4 presents a plot of the probability of a correct response and total test score. As shown in Figure 4, the

probability of getting item 2 correct is higher for the students who used a calculator to answer the question (Form D).

For item 1, testtakers need to know the meaning of 'average' and apply the property of average to restructure the story problem into a solvable form.(attributes 1, 5 6) (See Appendix A for the text of item #1.) Item 2 is a geometry problem. Examinees need to know the meanings of slope and intercept and how to add two factors (attributes 1 and 4). Perhaps calculators use helped students to avoid arithmetic errors in calculating the addition of two factors.

----Insert Figures 3 and 4 here----

DISCUSSION

This purpose of this study was to investigate whether the MMPE test items intended to be 'calculator neutral' perform as expected and whether these kinds of items are neutral for both males and females. The findings of this study suggest that calculator neutral items can be constructed. Futhermore, these items were largely free from gender DIF. However a more interesting question, whether the items are differentially more difficult or easier for females depending on the use of a calculator was not investigated, because of sample size requirements when using logistic regression The study has several other limitations also.

The logistic regression approach is an a useful additional to studying DIF. The flexibility of this approach which provides the opportunity to investigate DIF in relationship to variables like ability, calculator use, and gender in combination is a distinct advantage. However, there is a cost with this flexibility, especially for smaller testing programs. Findings from simulation studies suggest with sample sizes of 250 per group, DIF is detected with 75% accuracy when using logistic regression (Swaminathan & Rogers, 1990) To attain 100% accuracy, sample sizes of 500

for each group are required. To investigate the question of different conditioning variables like calculator use, achievement, and gender simultaneously, substantially larger sample sizes than those available in this investigation were essential.

Second, the pilot test items were clearly multidimensional in at least two respects: content, and problem type. Given that current analytic methods available for exploring DIF assume that the item responses are large unidimensional, interpretation of results for tests of this type are troublesome. For example items 9 and 1 which were differentially easier for males were story problems. The test specifications require story problem type items and test questions which are not story problems. However there will not be enough items of each type to be treated as a subtest. Stout and his colleagues are expecting to release a version of SIBTEST specifically designed to investigate DIF or differential bundle functioning (bundles of logically-linked items which do not function as intended) for tests designed to be multidimensional. This should be especially useful for examining tests like the one used in this investigation.

Using the rule-space approach as a basis for the content analyses was quite effective. The items identified as DIF in this investigation can be linked to other research on math and gender. Linn and Kessel (1994) also suggest males have better test-taking skills than females (item 7 Form AC) and that these differences may contribute to gender differences in math test scores. This investigation found that females find story problems differentially more difficult than males (items 9, Form AC and item 1, Form BD). There are similar findings from other gender DIF research with a similar population, SAT-M testtakers (Harris & Carlton, 1993). Results from other studies suggest females find both selection and supply type story problems especially difficult at a relatively early age (13 years) (Lane, Wang, & Magone, 1995, Ryan & Fan, 1994).

Both logistic regression and the rule-space approach basis to content analyses can contribute to further understanding of gender and math performance. In combination, these approaches may provide the opportunity tease out effects that may be linked to testing and teaching practices in mathematics (Linn & Kessel, 1995). This would be a major step forward: the possibility of understanding DIF, not just identify it

Reference

Ansley, T. N., Spratt, K. F., & Forsyth, R. A. (1989). The effects of using calculators to reduce the computational burden on a standardized test of mathematics problem solving Educational and Psychological Measurement, 49, 277-286.

Bridgman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning, Journal of Educational Measurement, 32, 4, 323-40

Camilli, G. & Shepard, L. (1994). Methods for identifying biased test items, Sage: Thousand Oaks, CA.

Cohen, A S., & Kim, S. (1992). Detecting calculator effects on item performance. Applied Measurement in Education, 5, 303-320.

Harnisch, D L., Tatsuoka, K.K., Wilkins, J L. (1995, November) Reporting math proficiencies based on new SAT-M items, Paper presented at the 1995 American Evaluation Association Meeting, Vancouver, BC.

Harvey, A. L., Jackson, C. P , & Faecher, K. E (1993) Findings from the first administration of the ATP mathematics level 2 achievement test with calculators ETS Report.

Harris, A. M. & Carlton, S. T (1993) Patterns of gender differences on mathematics items on the Scholastic Aptitude Test, Applied Measurement in Education, 6(2), 137-151

Lane, S , Wang, N., & Magone, M. (1995). Gender-related Differential item functioning on a middle school mathematics performance assessment, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Linn, M., & Kessel, C. (1995, April). Grades or scores: Predicting college mathematics performance. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Loyd. B H (1991) Mathematics test performance The effects of item type and calculator use Applied Measurement in Education, 4, 11-22.

Nandakamur, R (1993) Simultaneous DIF amplification and cancellation. Shealy-Stout's test for DIF, Journal of Educational Measurement, 30, 293-312.

National Council of Teachers of Mathematics. (1989) Curriculum and evaluation standards for school mathematics Reston, VA Author

Ryan, K. & Fan, M (1994, April). Gender differences on a test of mathematics: Multidimensionality or differential test functioning, Paper presented at the American Educational Research Association Annual Meeting, Atlanta, GA..

Ryan, K. & Fan, M. Validation studies for UIUC German, Spanish, and French Placement Examinations: Form 2", Measurement and Evaluation Division, Office of Instructional Resources, University of Illinois, Urbana, IL, Technical report prepared for American College Testing (ACT), Iowa City, IA under contract with the University of Illinois, (July, 1993).

Swaminathan, H , & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370

Tatsuoka, K.K (1993). Item construction and psychometric models appropriate for constructed response models. In R. Bennett & W. Ward (Eds.), Construction versus Choice in Cognitive Measurement, 107-133. Hilldale, NJ: Erlbaum

Wainer. H & Steinberg, L.S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study, Harvard Educational Review, 62, 323-326

Table 1

DIF Analysis Design

Analyses	Model	Groups Compared	DIF Type
Set 1	Full model Item = Score + Gender + Score*Gender Score + CNC + Score*CNC	Males / Females and achievement; C/NC and achievement;	Non-uniform
Set 2	Simpler model Item = Score + Gender Score + CNC	Males/ Females .C/NC	Uniform

Note. CNC means C=calculator allowed;NC= no calculator use allowed.

Table 2

Descriptive Statistics for MMPE Pilot Test

Group	N	Mean	Std Dev	Min	Max
Forms A and C					
Males	303	7 57	3.09	0	16 00
Females	262	6 68	2.89	0	16.00
No calculator	316	6 97	2 95	0	15.00
Calculator	249	7.41	3 12	1 00	16.00
Overall	565	7 16	3.03	0	16.00
Forms B and D					
Males	297	4 54	2.45	0	12.00
Females	254	3 67	2.12	0	12.00
No calculator	307	3 97	2.27	0	11.00
Calculator	244	4.35	2 42	0	12.00
Overall	551	4 14	2.34	0	12.00

Table 3

Reliability Analyses for the MMPE Pilot Test: KR-20 and Spearman-Brown Prophecy

Form	KR-20	Spearman-Brown Prophecy 40 items	Spearman-Brown Prophecy 45 items
A and C	0.648	0.812	0.829
A	0.629	0.800	0.818
C	0.667	0.825	0.842
B and D	0.608	0.838	0.854
B	0.583	0.823	0.840
D	0.635	0.853	0.867

Table 4

Summary of Logistic Regression Analysis for Set 2 Analyses

Form	Item	Effect	P-value	Parameter Estimate	Favors
A&C	7	Gender	0.016	-0.468	Male
	9	Gender	0.002	-0.583	Male
B&D	1	Gender	0 001	-0.621	Male
	2	CNC	0.030	-0 486	Calculator

Note. CNC means calculator allowed and no calculator allowed. For gender. females are the focal group and males are the reference group. For CNC, the group that used the calculator is the reference groups; the focal group did not use a calculator.

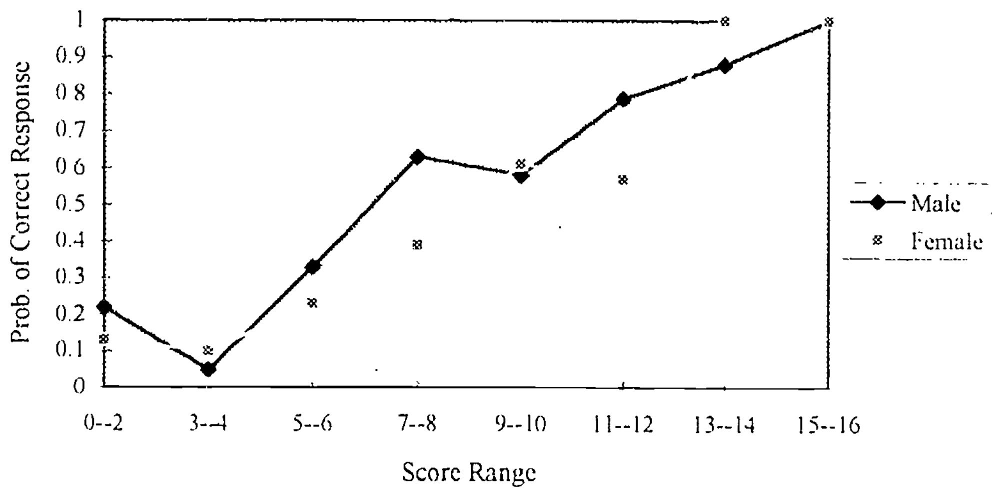

Figure 1 Plot of the probability of a correct response for item 7 (Forms A and C) and total score for males and females.

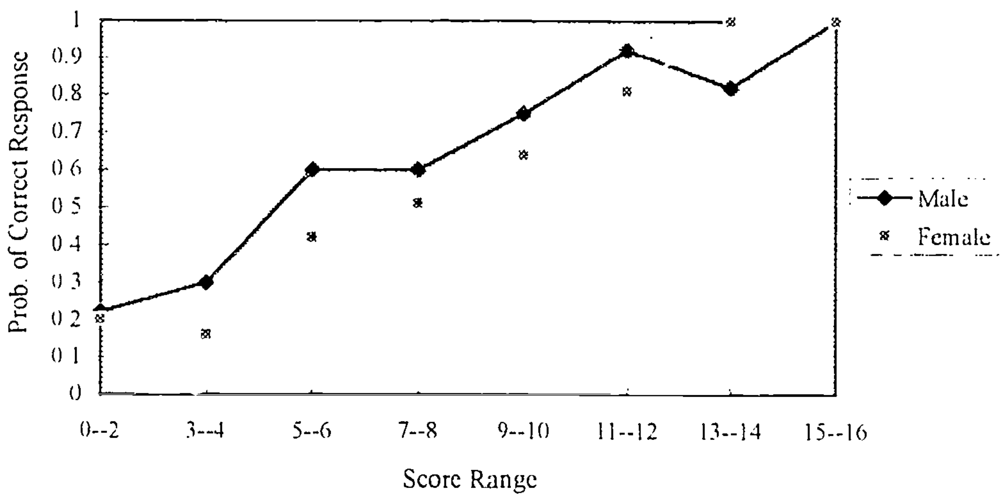

Figure 2. Plot of the probability of a correct response for item 9 (Forms A and C) and total score for males and females.

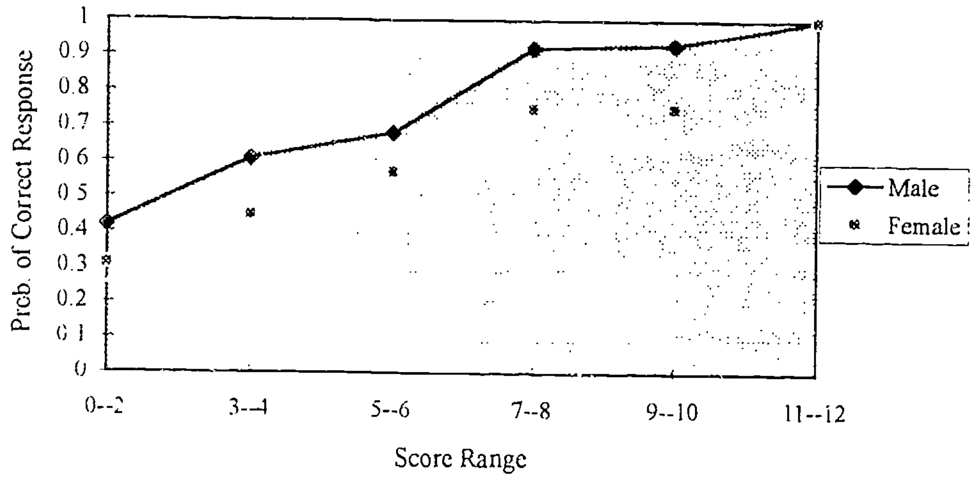

Figure 3: Plot of the probability of a correct response for item 1 (Forms B and D) and total score for males and females.

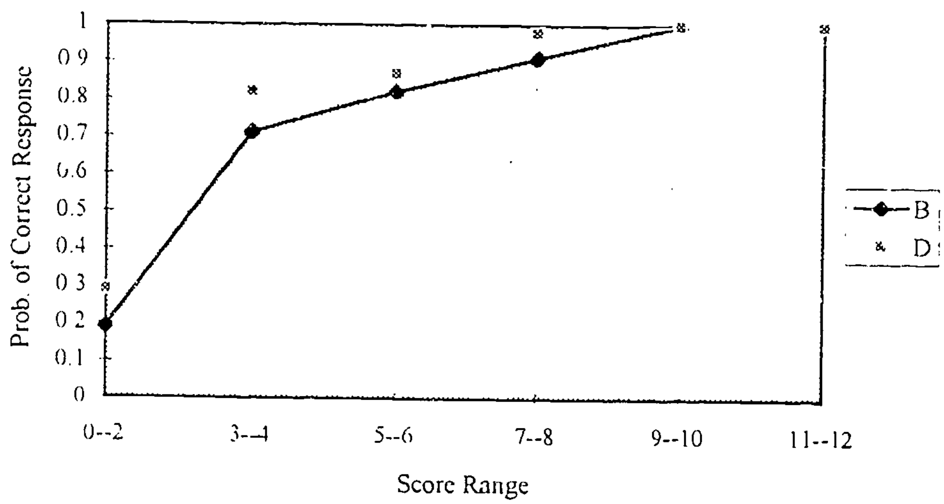

Figure 4: Plot of the probability of a correct response for item 2 and total score for Form B and Form D testtakers.

21

APPENDIX A

Forms A and C

7 The smallest possible value of $f(x) = (x-5)^2 - 5$ is

(a) -10 (b) -5 (c) 0 (d) 20 (e) 5

Forms B and D

1. Sam received grades of 87, 75, and 72 on three math tests. What average does he need on the next two tests in order to average 80 on all five?

(a) 81 (b) 82 (c) 83 (d) 85 (e) 86

Description of Mathematical Challenges:

1 - Dealing with odd & even integers, prime numbers, factors, rational numbers, ordering, ratios, percentages, place value, powers, roots, and averages.

2 - Dealing with variable (addition and subtraction only), linear equations, linear algebraic expressions, signed-numbers, absolute values, irrational numbers.

3 - Dealing with higher-degree algebraic expressions, functions,sets, simple probability, combinatorics, modes and medians, exponents with variables.

4 - Dealing with perimeter, area and volume for triangles, circles, rectangles and other geometric objects. In analytic geometry, dealing with points, lines, in relation to a coordinate system.

5 - Translating word problems into arithmetic and algebraic expression(s). Can identify implicit variables and relationships. Dealing with real-world problems and real-world experiences.

6 - Restructuring problems into solvable forms. Choosing better, simpler or quicker strategies to solve problems. Choosing from rules, properties and theorems the better, simpler or quickest one to use.

7 - Recalling and interpreting knowledge based on definitions, properties or relationships from arithmetic, algebra, and geometry. Can perform computations in arithmetic, geometry, signed numbers, absolute value, median and mode.

8 - Applying mathematical rules and properties to solve equations (simultaneous); derive, factor and compute algebraic expressions.

9 - Reasoning and logical thinking. Reasoning deductively from cause to effect. Spatial reasoning skills. Identify and understand necessary and sufficient conditions and apply them.

10 - Application of higher mental processes to solve problems. Sorting problems into implicit component parts and restructuring them in order to make the problem solvable.

11 - Working with figures, tables, and graphs. Can generate figures to facilitate problem-solving activities.

12 - Can take advantage of the form of the test items and other test-taking methods without solving the problem in the manner intended by the item writer. Can solve a task by working backwards from the multiple-choice options.

13 - Working with problems having several steps. These steps may be explicit or implicit. Can establish subgoals of the problem, order, prioritize and execute the subgoals in a step-by-step fashion.

14 - Can comprehend sentences with negation, "at least", comparison, "must be", "could be", and with the relations of increasing and decreasing.

15 - Answering questions formatted as grid-ins. Deriving solutions by a top down approach.

16 - Keeping track of what a question is asking, paying attention to detail. Identify constraints. Follow verbally written instructions, read complex, long sentences.

17 - Straight forward translations of verbal expressions into mathematical expressions where variable term(s), constant(s), and needed operation(s) are readily apparent.

From Harnisch, D., Tatsuoka, K.K., & Wilkins, J.L. (1995, November). Reporting math proficiences based on new SAT-M items, Paper presented at the 1995 American Evaluation Association Meeting, Vancouver, BC